Bootstrap

Marcin Kostrzewa

7 kwietnia 2022

Uwagi wstępne

Najpierw rozróżnijmy podejście parametryczne od nieparametrycznego. Polegają one kolejno na:

- *w przypadku podejścia parametrycznego* — zakładamy w naszym modelu znajomość rozkładu, z którego pochodzi analizowana próbka,
- *w przypadku podejścia nieparametrycznego* — nie zakładamy znajomości rozkładu, z którego pochodzi posiadana próbka.

Po co w ogóle bootstrap?

Zakładając znajomość rozkładu próby jesteśmy w stanie (choć nie zawsze, a jak już, to nie jest to łatwe) wyznaczyć rozkład statystyki estymującej daną cechę, podać przedziały ufności, zastosować test statystyczny. Co jednak, jeśli nie znamy rozkładu, a mamy do dyspozycji jedynie pewną próbkę? Z jednej wartości liczbowej trudno wyciągać jakiekolwiek wnioski, trzeba zaproponwoać inne podejście.

Co to? (Bootstrap nieparametryczny — pomysł Efroniego)

Metoda bootstrap polega na wylosowywaniu z próbki, którą posiadamy, B (domyślnie liczba pokaźnych rozmiarów) próbek tej samej wielkości. Losowanie odbywa się z powtórzeniami. Tak stworzone próbki nazywamy replikacjami bootstrapowymi lub po prostu próbkami bootstrapowymi.

Schemat postępowania

Załóżmy, że mamy próbkę $X_1, X_2, \ldots, X_n$. Interesuje nas oszacowanie i przeprowadzenie wnioskowania dla cechy θ.

1. Uzbrajamy się w statystykę $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$.
2. Powtarzamy B-krotnie — dla $i = 1, \ldots, B$:
   2.1 tworzymy replikację bootstrapową $X_1^{(i)}, \ldots X_n^{(i)}$,
   2.2 wyznaczamy $\hat{\theta}_i^* = T\left(X_1^{(i)}, \ldots, X_n^{(i)}\right)$.
3. Wnioskujemy na podstawie uzyskanych oszacowań $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

Wyznaczenie błędu

Co z wariancją $\hat{\theta}$? W bootstrapie możemy dobrze przybliżyć ją za pomocą wariancji replikacji bootstrapowych, czyli za pomocą:

$$\mathrm{Var}(\hat{\theta}^*) = \frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2,$$

gdzie $\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$.

Błąd standardowy estymatora to $\mathrm{SE}\left(\hat{\theta}\right) \approx \sqrt{\mathrm{Var}(\hat{\theta}^*)}$.

Konstrukcja przedziałów ufności (1)

Wyróżnia się różne metody konstrukcji przybliżonych przedziałów ufności. Załóżmy, że dzięki metodzie bootstrap uzyskaliśmy zbiór replikacji bootstrapowych $\hat{\Theta}^* = \left(\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\right)$.

- Przedziały kwantylowe — bierzemy po prostu kwantyle odpowiedniego rzędu z $\hat{\Theta}^*$, uzyskując przedział:

$$\left(Q_{\frac{\alpha}{2}}\left(\hat{\Theta}^*\right), Q_{1-\frac{\alpha}{2}}\left(\hat{\Theta}^*\right)\right)$$

- Przedziały BBCI (*bootstrap basic confidence interval* — nikt poza mną nie używa takiego skrótu) — mają one postać następującą:

$$\left(2\hat{\theta} - Q_{1-\frac{\alpha}{2}}\left(\hat{\Theta}^*\right), 2\hat{\theta} + Q_{\frac{\alpha}{2}}\left(\hat{\Theta}^*\right)\right)$$

- Przedziały „normalne” — zakłada się normalność $\hat{\theta}^*$.

Konstrukcja przedziałów ufności (2)

- Przedzialy studentyzowane — zakłada się podobieństwo rozkładu:

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \approx \frac{\hat{\theta}^* - \hat{\theta}}{\sigma_{\hat{\theta}^*}} = q.$$

  Przedziały przyjmują postać:

$$\left(\hat{\theta} - \sigma_{\hat{\theta}} Q_{1-\frac{\alpha}{2}}\left(\mathbf{q}\right), \hat{\theta} + \sigma_{\hat{\theta}} Q_{\frac{\alpha}{2}}\left(\mathbf{q}\right)\right).$$

  $\sigma_{\hat{\theta}} \approx \mathrm{SE}\left(\hat{\Theta}^*\right)$, $\mathbf{q} = (q_1, \ldots, q_B)$, a dla każdej próbki bootstrapowej $\sigma_{\hat{\theta}^*}$ wyznaczamy za pomocą …metody bootstrap.

Konstrukcja przedziałów ufności (3)

- Przedziały BCa (ang. *bias-corrected*), które przybierają postać:

$$\left(\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2}\right), \quad \text{gdzie}$$
$$\alpha_1 = \phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_\alpha}{1 - \hat{a}\left(\hat{z}_0 + z_\alpha\right)}\right)$$
$$\alpha_2 = \phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha}}{1 - \hat{a}\left(\hat{z}_0 + z_{1-\alpha}\right)}\right),$$

gdzie $\hat{a}$, $\hat{z}_0$ to pewne parametry, z_α to kwantyl rozkładu normalnego, a $\phi(\cdot)$ to jego dystrybuanta

Czas na przykład

Z lekka więcej teorii …

Do tej pory powiedzieliśmy sobie tylko, jak korzystać z bootstrapu, jak wyznaczać błąd standardowy czy przedziały ufności. Na pewno w niektórych z Waszych głów pojawiło się pytanie — co uzasadnia korzystanie z tej metody?

Estymatory typu plug-in

Estymatory możemy wyznaczać jako funkcjonały — przekształcenie z przestrzeni funkcji w pewne ciało (np. $\mathbb{R}$). Najczęściej będziemy mieli do czynienia z funkcjonałami dystrybuany.

Przykład — średnia:

$$\mu = T(F) = \int x \, \mathrm{d}F(x)$$

Istnieje taki twór jak dystrybuanta empiryczna — $\hat{F}_n(x)$. Jeżeli podstawimy sobie w miejsce F dystrybuantę empiryczną $\hat{F}_n$ otrzymamy estymator pewnej cechy statystycznej — $\hat{\theta} = T(\hat{F}_n)$, to otrzymamy estymator typu *plug-in*.

Czemu bootstrap działa?

Sama empiryczna dystrybuanta to porządny estymator — nieobciążony o wariancji zbiegającej do 0, zgodny.

Tworzenie próbki bootstrapowej to tak naprawdę generowanie próbki z rozkładu zadanego przez empiryczną dystrybuantę.

Zatem replikacje bootstrapowe można uznać za estymacje za pomocą estymatora plug-in $\hat{\theta} = T\left(\hat{F}_n, n\right)$, który najczęściej dobrze przybliża estymator dokładny $\hat{\theta} = T\left(F, n\right)$.

Inne rodzaje bootstrapu

- Bootstrap parametryczny
- Bootstrap wygładzony (*smoothed*)
- Booootstrap wpół-parametryczny (*semiparametric*)
- Bootstrap bayesowski

Gdzie jeszcze używamy bootstrapu?

- Bagging
- Regresja liniowa
- Szeregi czasowe

Podsumowanie

Kiedy warto sięgnąć?

- małe próbki (problem z asymptotycznością),
- skomplikowana estymacja.

Kiedy nie stosować?

- Statystyka T nie jest dostatecznie gładka.
- Próbka jest bardzo mała.
- Zajrzyj tutaj.

Dzięki za uwagę :)

Źródełka

- Efron B., Tibshirani R.J., An Introduction to the Bootstrap
- Yen-Chi Chen, The Bootstrap
- Jacek Gulgowski, Barbara Wolnik, Metody bootstrapowe w statystyce
- Taki wykładzik